NSW GOVERNMENT | Department of Primary Industries

Final Report

Genome sequencing of myrtle rust, *Puccinia psidii* sensu lato

Mui-Keng Tan, Damian Collins, Zhiliang Chen, Anna Englezou and Marc Wilkins

30 June 2013

A report for Plant Health Australia Ltd

Project Leader contact details:

Dr. Mui-Keng Tan
Biosecurity NSW, NSW Department of Primary Industries,
Elizabeth Macarthur Agricultural Institute, Woodbridge Road, Menangle,
NSW 2568
+61246406233
+61246406234
mui-keng.tan@dpi.nsw.gov.au

ACKNOWLEDGEMENTS

- Professor Robert Park, director of Cereal Rust Research, PBI, University of Sydney and his staff, Keshab Candle and Dr. Karanjeet Sandhu for the supply of a large quantity of urediospores of a single isolate of myrtle rust pathogen, 115012-Mr for large scale extraction of DNA;

- Dr. Indu Sharma (Punjab Agricultural University) and Dr. Harsh Raman (NSW DPI) for the provision of a DNA sample of an isolate of *Tilletia indica*.

- We thank Plant Health Australia and Department of Agriculture, Fisheries and Forestry (DAFF), Canberra, Australia for financial support through the Myrtle rust Transition to Management research response program.

Table of contents

1. Executive Summary

Using *De novo* assembly of 46 million paired end sequence reads of length 250 bp for a myrtle rust isolate, we have estimated the genome size to be between 103 –145 Mb and the number of proteins as >19,000. Mapping and annotation of the contig sequences found a very large percentage of proteins are associated with molecular functions of DNA binding or binding in biological processes for DNA integration and RNA-dependent DNA replication. A large proportion of these activities are attributed to the transposable elements (TEs). These elements are estimated to comprise 27% of the genome with 22 % retrotransposons and 5% DNA transposons. They are thus the major players in the generation of genetic variability in the pathogen's adaptation to the environment and new hosts. It is thus postulated that myrtle rust is a complex of genotypes from races to species and possess the genetic ability to jump across hosts of different species and genera of *Myrtaceae*. The exon and intron boundaries of forty six genes occurring on contigs > 20,000 bp have been determined. The β-tubulin gene has been annotated on a contig of length 3,439 bp. All these myrtle rust gene sequences have been submitted to GenBank. The number of introns range from 2 to 20 with a mean of 7.3. Phylogenetic analysis using the clathrin associated protein, the β-tubulin gene and the partial *COXI* gene from this work has placed myrtle rust in the major clade of *Pucciniaceae* in the *Pucciniales* lineage but on a separate taxonomic branch from the one for the cereal rust fungi and *Phakopsora* species.

2. Introduction and Objectives

Myrtle rust (*Puccinia psidii* sensu lato) was reported for the first time in Australia in April 2010 from *Agonis flexuosa*, *Callistemon viminalis* and *Syncarpia glomulifera* (Carnegie et al. 2010). These rusts are serious pathogens which affect plants belonging to the family *Myrtaceae* including Australian natives like bottle brush (*Callistemon* spp.), tea tree (*Melaleuca* spp.) and eucalypts (*Eucalyptus* spp.). Since its first report, the disease is currently found to be widely distributed along the entire east coast of New South Wales and parts of Queensland and Victoria.

Myrtle rust produces masses of powdery bright yellow or orange-yellow spores on infected plant parts (Fig. 1). It infects susceptible plants producing spore-filled lesions on young actively growing leaves, shoots, flower buds and fruits. Leaves may become buckled or twisted and may die as a result of infection. Sometimes these infected spots are surrounded by a purple ring. Older lesions may contain dark brown spores. Infection on highly susceptible plants may result in plant death.

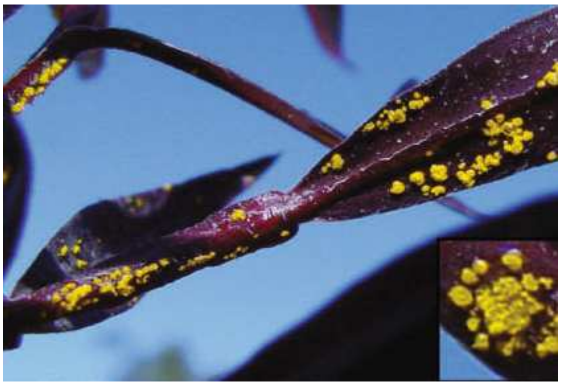

Figure 1. Masses of bright yellow uredinia on leaves of *Agonis flexuosa* cv. 'Afterdark' (Carnegie et al. 2010).

Myrtle rust is morphologically distinct from eucalyptus/guava rust (*Puccinia psidii*), and was initially identified as *Uredo rangelii* (Carnegie et al. 2010). The subsequent discovery of teliospores from myrtle rust which matched those of *P. psidii* sensu stricto led to its name being revised to *P. psidii* sensu lato (Carnegie and Cooper 2011). The pathogen is referred in this report as myrtle rust to distinguish it from eucalyptus/guava rust.

Guava rust was first reported on a native South American *Myrtaceae*, *Psidium pomiferum* (Winter 1884). *P. psidii* sensu lato has since been reported in majority of countries in South and Central America, Florida, California and Hawaii in the United States (Coutinho et al. 1998, Glen et al. 2007). Rusts occurring on *Myrtaceae* usually has a wide variety of hosts but *P. psidii* Winter is the only myrtle rust that is capable of infecting *Eucalyptus* species (Junghans et al. 2003). This rust was similarly reported to have a wide host range across genera and very damaging to eucalypt plantations in Brazil and considered a serious threat to eucalypt plantations worldwide (Coutinho et al. 1998). Some of these *Eucalyptus* species introduced for commercial purposes originated from Australia.

Myrtle rust has to date been documented on 107 host species in 30 genera from data collected during the 2010 surveys in NSW under the state emergency response program (Carnegie and Lidbetter 2012). Host range studies performed using artificial inoculation experiments (Carnegie and Lidbetter 2012) have showed that several species of Australian *Eucalyptus* are susceptible to myrtle rust. There is to date no report of a natural infection on *Eucalyptus* in Australia.

The threats myrtle rust poses to the Australian flora and the forestry industry worldwide make it crucial to have an in-depth genetic understanding of the fungus. There is to date very limited genetic data available for Eucalyptus/guava rusts and they were thus unable to shed any light on the myrtle rust fungus.

A project was proposed in Aug-2012 to use next generation sequencing technologies to obtain genome sequence data of the type isolate of myrtle rust (115012-Mr). This will enable the genomics data to complement conventional taxonomic work involving morphology and classical sequencing to determine the taxonomic status of the pathogen in the huge complex of rust fungi.

The objective of this project was to provide some insights into the myrtle rust genome from whole genome sequencing of the pathogen and to use appropriate gene regions to determine the taxonomic status of the pathogen. This knowledge will contribute to clarify the current confusion on the taxonomic status of *P. psidii* (van der Merwe et al. 2008) and enhance the genetic understanding of the pathogen for the development and implementation of long term management strategies of the pathogen.

3. Methods

3.1. DNA extraction and Genome Sequencing

A single urediospore of the myrtle rust pathogen, PBI accession no, 115012-Mr was inoculated on the host, *Syzygium jambos* (rose apple) in August 2012. Multiplication of rust spores was performed at PBI, University of Sydney. About 0.1 g of spores was ground in extraction buffer (50 mM Tris-HCl [pH 8.0], 0.7 M NaCl, 10 mM EDTA, 1% [wt/vol] cetyl trimethylammonium bromide [CTAB; Sigma H-5882], and 1% [vol/vol] 2-mercaptoethanol) for DNA extraction. High quality, un-degraded DNA was extracted as outlined in Tan and Niessen 2003.

A shot-gun library of sequences of about 650 bp was prepared using the TruSeq DNA Sample Preparation kit (http://www.illumina.com/). The library of random DNA fragments from the entire genome was sequenced on the MiSeq Sequencer (http://www.illumina.com/systems/miseq.ilmn) at the Ramaciotti Centre, University of NSW. Two separate sequencing runs were performed from the library.

A similar sequencing run on the MiSeq Sequencer was performed on a DNA sample from an isolate of *Tilletia indica* Ps23 isolated from host *Triticum aestivum* in 1990 at Gurdaspur, India. The sequence data were used to compare the outcomes of *de novo* assemblies with myrtle rust data in the discussion. Analysis of genome sequence data for *T. indica* will not be presented here.

3.2. Data Analysis-Sequence Quality

A program called FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to check the quality of the reads.

The combined sequences from both runs were trimmed using the 'trim' tool in the CLC Genomics Workbench 6 (www.clcbio.com). A limit value of 0.05 for the quality trimming with a maximum number of two ambiguous nucleotides at the sequence ends allowed.

3.3. Sequence Assembly

There are two commonly used algorithms for assembly: the overlap graph and the de Bruijn graph. Fast and memory efficient algorithms based on the computation of suffix-prefix matches among all pairs of reads in open source softwares, Readjoiner (Gonnella and Kurtz 2012) and the String Graph Assembler (SGA) (Simpson and Durbin 2011) were first used to assemble the data. The programs were reiterated with different overlap lengths (*ol*) to obtain the optimum assembly with the longest contig and highest N50 value.

The open source software, ABySS (Simpson et al. 2009), the SOAPdenovo (http://soap.genomics.org.cn/) and the commercial software, CLC-Bio

Genomics Workbench were based on the de Bruijn algorithm and were also used to assemble the data. Assembly was reiterated with different k-values to obtain the best output in terms of the length and number of contigs and N50 value.

3.4. Analysis, Mapping and Annotation of contigs

De novo assembly data used for analysis was the set of contigs with length, *l* ≥ 3 kb generated by the assembly program, CLC-Genomics Workbench which gave the highest N50 value (Table 2). Contigs in the range 3 to 8 kb were blasted, mapped and annotated using the Blast2Go software (http://www.blast2go.com/b2ghome).

Contigs longer than 8 kb were blasted using the Blastx program on the NCBI site (http://blast.ncbi.nlm.nih.gov/Blast.chi). The results were imported into a Blast2Go file for mapping and annotation.

All blastx analysis were ran against the fungal set of the non-redundant protein sequences, with a word size of 3, expectation value of 10.0 and the number of blast hits archived was limited to 5. The scoring parameters were BLOSUM62, Existence:11, Extension:1, with conditional compositional score matrix adjustment.

The blast results (blast result accessions) were mapped using Blast2GO to retrieve Gene Ontology (GO) terms associated with the hits. The mapping step was followed by GO annotation using the default parameters in the Blast2Go program. The annotated sequences were then analysed using the data mining tool in Blast2Go with respect to the distribution of the annotated sequences in the cellular component, molecular function and biological process of the genome (Götz et al. 2008).

3.5. Determination of proportions of repeats

Computations were performed in the R statistical language to calculate the proportions of repeats in the rust genome. Text strings for repeated sequences were used to match against the 'sequence description' of the Blast2Go output. They were retrotransposable elements (pol, gag, rve, integrase, polyprotein, retrotransposable, retrotransposon, tick, reverse transcriptase, RNAse H); copia polyprotein (Ty1, copia); gypsy retrotransposon (Ty3, gypsy, nucleocapsid); DNA transposons (Tc1, hAT, DDE, mutator, transposase); hAT; mutator; Tc1.

3.6. Fine annotation of protein genes

Blast results with significant homology (E-value < -50) with closely related gene sequences from other fungal organisms e.g. *P. graminis* were used as starting points to perform fine annotation of genes on the set of contigs with length >20,000 bp. Each of the query contig from myrtle rust with a highly significant blastx result for a protein gene will have multiple segments of the contig matched to segments from the five archived orthologous gene sequences with different homology values (E- value). The nucleotide interval of the query contig that had the highest homology (=lowest E-value <-50) was used as an anchor point to determine the putative boundary of that exon.

Using the blast result as a guide, a nucleotide interval downstream or upstream of the putative boundary of that exon was then selected. This selection was translated in 3 different frames and aligned with corresponding protein segments of gene orthologs. The frame that gave the highest homology was used to determine the next exon. The process was repeated until all exon sequences of the gene have been determined.

3.7. Phylogenetic Analysis

Phylogenetic analysis of protein and DNA sequences was performed using the program PAUP* Version 4b10 (http://paup.csit.fsu.edu/). In the DNA analysis, the codon position was defined and the transition to transversion weighting ratio of 1 to 3 was used.

4. Results

4.1. Genome Sequencing

A sufficient quantity of rust spores for large scale DNA extraction was obtained by the inoculation of a single urediniospore of myrtle rust isolate, 115012-Mr on the host, *Syzygium jambos*. Genomic DNA of the pathogen extracted using commercial kits from Viogene (http://www.viogene.com/) and Qiagen (www.qiagen.com) failed the QA specifications for DNA quality for library preparation for next generation sequencing. Un-degraded DNA of high purity suitable for library preparation was obtained using the CTAB method (Tan and Niessen 2003).

Two separate NGS runs were performed on the MiSeq Sequencer (http://www.illumina.com/systems/miseq.ilmn), which generated 18 million and 28 million paired reads of 250 bp per read. A total of 46 million paired reads of length 250 bp were thus obtained from two sequencing runs.

4.2. Sequence Quality

The program, FastQC, generates quality scores on the Phred scale for each position of the 250 bp-reads in the NGS data (Fig. 2). Phred score gives an indication of the quality by linking to the log of the error probabilities. Hence a Phred score of 10 gives an error probability of 1 in 10 and a base call accuracy of 90%. The reads in the library have Phred scores above 30 for positions from 1 to 200, suggesting an accuracy of >99.9 %. Reads beyond 250 dipped slightly in quality. The reads of both runs have a mean quality score of 37 giving a sequence error of less than 1 in 1000 and an accuracy of more than 99.9%.

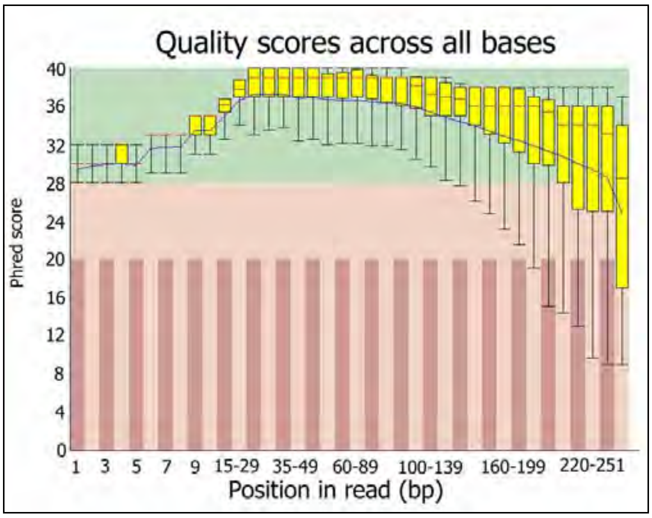

Figure 2. Per Base sequence quality of the read sequences on the Phred scale.

The trimming of the combined sequences from both runs reduced the average length from 250 bp to 242 and 234 bp for Run 1 and Run 2 respectively (Table 1, Fig. 3). It also removed ambiguous reads.

Table 1. Summary of the trim function on the raw sequence reads from two NGS run.

Name	# of paired end reads	Avg. length	# of paired end reads after trim	% trimmed	Avg. length after trim
Run 1	18,240,318	250	18,187,460	99.71	242.7
Run 2	28,285,366	250	28,243,206	99.85	234.1

Table 1. (continued)

Trim	Input reads	No trim	Trimmed	Discarded
Trim on quality	46,322,773	21,362,849	24,953,737	6187
Ambiguity trim	46,283,547	45,910,751	283975	88821

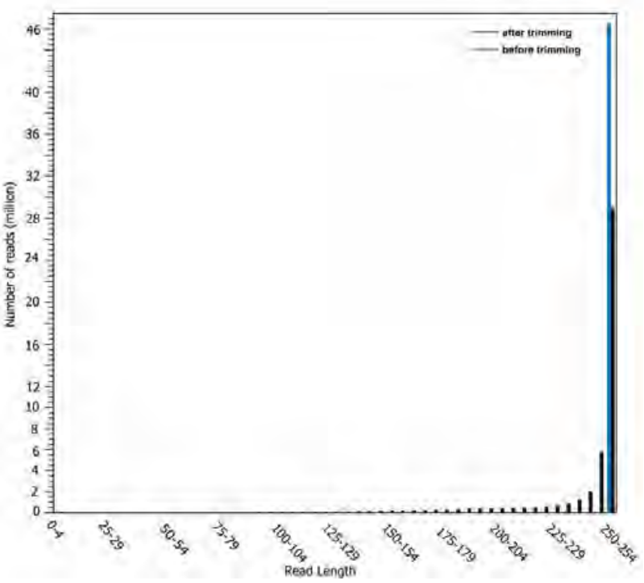

Figure 3. The distribution of the read lengths before and after the trim function.

4.3. Sequence Assembly and Estimation of Genome Size

The genome has about 15-16% G and 15-16% C which makes it about 30-32% GC rich. It has about 32-33% of A and T and thus about 64-66% AT content.

There is no reference genome for the myrtle rust pathogen. The genome has to be assembled *de novo* from the 46 million reads of 250 bp/read. A *de novo* DNA assembly is characterized by the N50 value, the maximum contig size and the number of contigs. An N50 contig size of N means that 50% of the assembled bases are contained in contigs of length N or larger. An optimum output is one with values as large as possible for the maximum contig size and N50 and as low as possible the number of contigs with the perfect number being the number of chromosomes (linkage groups) in the genome.

Assembly of trimmed sequence data using Readjoiner (Table S1) with different overlap lengths gave N50 values in the range 432–440 bp. The String Graph Assembler (SGA) gave a very similar N50 of 493 with an overlap length of 85 (Table 2).

DNA assemblers using the de Bruijn graph (Pevzner et al. 2001) reduces the computational charge by breaking reads into smaller sub-sequences of DNA, called k-mers, where the k parameter describes the length in bases of these sequences (Miller et al. 2010). The k-mer lengths used in the ABySS software and CLC-Bio Genomics Workbench were in the range 45–89 and 32–64 (longest length allowed is 64) respectively.

The ABySS software gave N50 values in the range 515–558 bp and maximum contig size from 20,332 to 35,730 bp (Table S2). The SOAPdenovo (http://soap.genomics.org.cn/) assembly program gave an assembly with N50= 567 bp and max contig length=6,094 bp for a high k-mer=127 (Table 2). These two "de Bruijn graph" assemblers seemed to perform a bit better than the two "string graph-based" assemblers on the N50 measure.

Table 2. A comparison of genome assembly of NGS data of myrtle rust using different *de novo* assemblers.

***De novo* assembler**		**Total # of contigs**	**N50 (bp)**	**Max Contig(bp)**	**Total length (bp)**
SGA ([1]*ol*=85)	all	1,321,742	493	8,631	573,387,137
	[2]*l* ≥750	135,453	976		137,671,501
Readjoiner (*ol*=70)	all	4,963,090	439	3,099	2,147,976,637
	l ≥750	116,914	859		103,828,908
AbySS (k=45)	all	759,095	558	25,668	362,000,000
	l ≥750	99,806	1,522		145,365,606
SOAPdenovo (k=127)	all	766,280	567	6,094	387,909,377
	l ≥750	109,692	949		108,060,557
CLC-Bio (k=40)	all	148,139	3,165	47,187	387,958,276
	l ≥750	148,138	3,165		387,956,878
	l ≥3000	37,684	5,535		203,507,520
	l ≥4000	23,106	7,929		153,506,636

[1] overlap length; [2] contig length (bp)

The CLC-Genomics workbench gave N50 values in the range 1946–3059 bp and the maximum contig sizes between 39,311 and 47,817 bp (Table S3). The N50 values from CLC-Genomics Workbench are much higher than from the other assemblers. A comparison of the outputs from the different assembly programs (Table 2) suggested the best output as one obtained using k-mer =40 on the CLC-Genomics Workbench.

The CLC-Genomics Workbench was unable to build longer contigs using the scaffolding function (Fig. 4) from the many primary contigs assembled, suggesting the presence of a large proportion of branching nodes due to ambiguities from repeated elements.

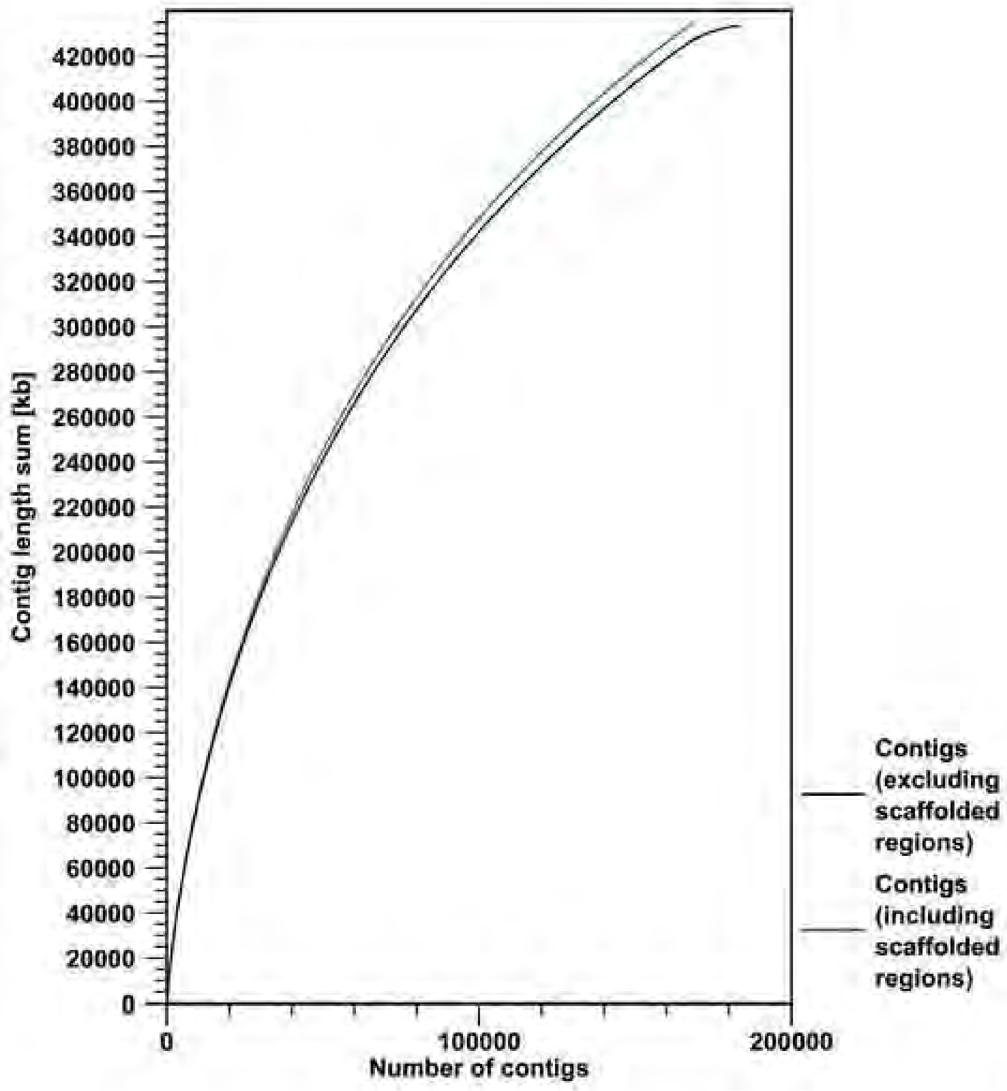

Figure 4. Comparison of contig length sum from contigs (excluding scaffolded regions) and contigs (including scaffolded regions).

The sets of contigs assembled using the various softwares (Table 2) were used to estimate the genome size. For assemblies with N50 values around 500 bp, the sets of contigs with threshold length of at least 750 bp were summed to give an estimate of genome size. Hence the genome size was estimated to range from 103 Mb (Readjoiner) to 145 Mb (ABySS).

4.4. Blast Analysis

The set of contigs ≥ 3 kb obtained from the CLC-Genomics Workbench (Table 2) was used in blastx analysis (http://blast.ncbi.nlm.nih.gov/BLAST) with non-redundant protein sequences to analyse the protein composition. Due to the huge number of contigs, the number of blast hits archived was limited to 5. The highest number of matches was obtained with the species, *P. graminis* (Fig. 5). This number was about 54,000. The second largest number of matches was about 6,000 and was with another rust species, *Melampsora larici-populina*.

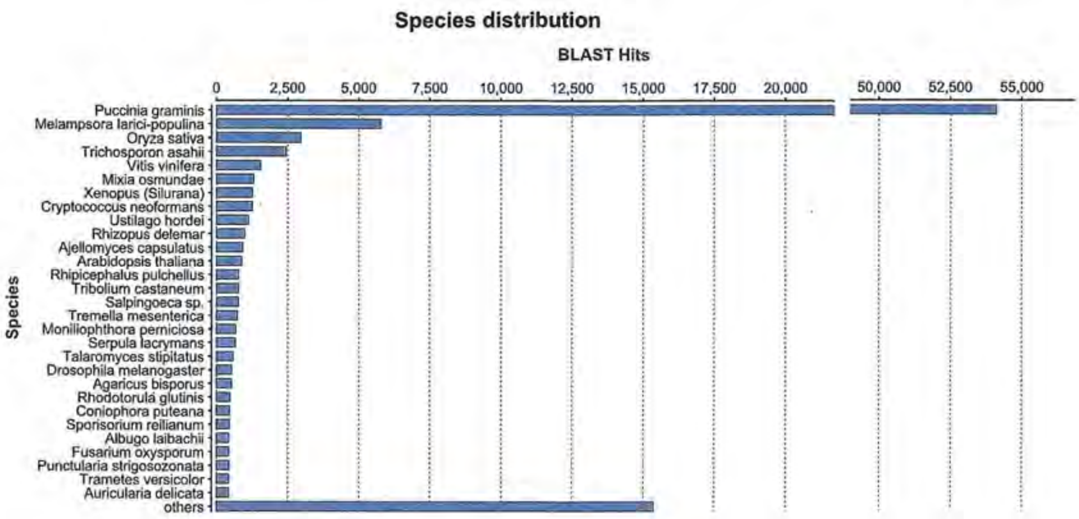

Figure 5. The distribution of species blastx hits with contigs of the myrtle rust genome.

The blast results (blast result accessions) were mapped using Blast2GO to retrieve Gene Ontology (GO) terms associated with the hits. The database resource for mapping the proteins was almost entirely from UnitProtKB with the evidence code for mapping derived overwhelmingly from electronic annotation.

Annotation is the process of selecting GO terms from the GO pool obtained by the mapping step and assigning them to the query sequences. The selection is based on the computation of an annotation score (http://www.blast2go.com/b2ghome). The contig sets, *l* ≥ 3 and *l* ≥ 4 kb from the CLC-Genomics Workbench resulted in 2,907 and 2,032 proteins annotated respectively using the default parameters in Blast2GO (Table 3).

The annotated sequences were analysed using the data-mining tool in the Blast2GO software with respect to the distribution of annotated proteins (GO-terms) in the cellular component, biological process and molecular function of the genome.

The highest number of GO counts in the cellular component was found for proteins in the nucleus, with significant counts from cytoplasm, membrane and mitochondrion. The two largest groups of GO terms in biological process are proteins for DNA integration and RNA-dependent DNA replication (Fig. 6). Proteins in these categories comprise largely the transposable elements. The GO terms in molecular function are mainly associated with binding, with the highest proportion involved in nucleic acid binding (Fig. 6). RNA binding and RNA-directed DNA polymerase activities are also significant indicating the activities of the retrotransposons.

4.5. Protein Composition

The protein composition of myrtle rust genome was estimated from the contig set, $l \geq 3$ kb and comprises ~10,105 hypothetical proteins, 6,800 mapped proteins and 2,907 annotated proteins (Table 3) giving the estimated number of proteins as 19,812 (Table 3).

Table 3. Protein composition in the myrtle rust genome

Proteins	Contig set $l \geq 3$ kb		Contig set $l \geq 4$ kb	
	# of contigs	%	# of contigs	%
Total # of contigs	37,605		23,106	
No blastx hits	16,246	43.2	8,489	36.7
mapped proteins	6,800	18.1	4,932	21.3
Blast2GO annotated proteins	2,907	7.7	2,032	8.8
Mitochondrial contigs	79	0.21	65	0.28
Hypothetical protein -[Uncharacterized protein hypothetical protein [1]PGTG_XXXXX hypothetical protein [2]MELLADRAFT_XXXXX hypothetical protein [3]TREMEDRAFT_XXXXX hypothetical protein [4]MPER_XXXXX Others]	10,105 [21 8,803 704 19 30 8]	26.9 [0.1 23.4 1.9 0.1 0.1 0]	6,927 [16 5,994 496 17 20 7]	30 [0.1 25.9 2.1 0.1 0.1 0]

[ ] indicates the breakdown of hypothetical protein.

[1] PGTG_XXXXX: hypothetical protein [*Puccinia graminis f. sp. tritici* CRL 75-36-700-3]

[2] MELLADRAFT_XXXXX: hypothetical protein [*Melampsora larici-populina* 98AG31]

[3] TREMEDRAFT_XXXXX: hypothetical protein [*Tremella mesenterica* DSM 1558]

[4] MPER_XXXXX: hypothetical protein [*Moniliophthora perniciosa* FA553]

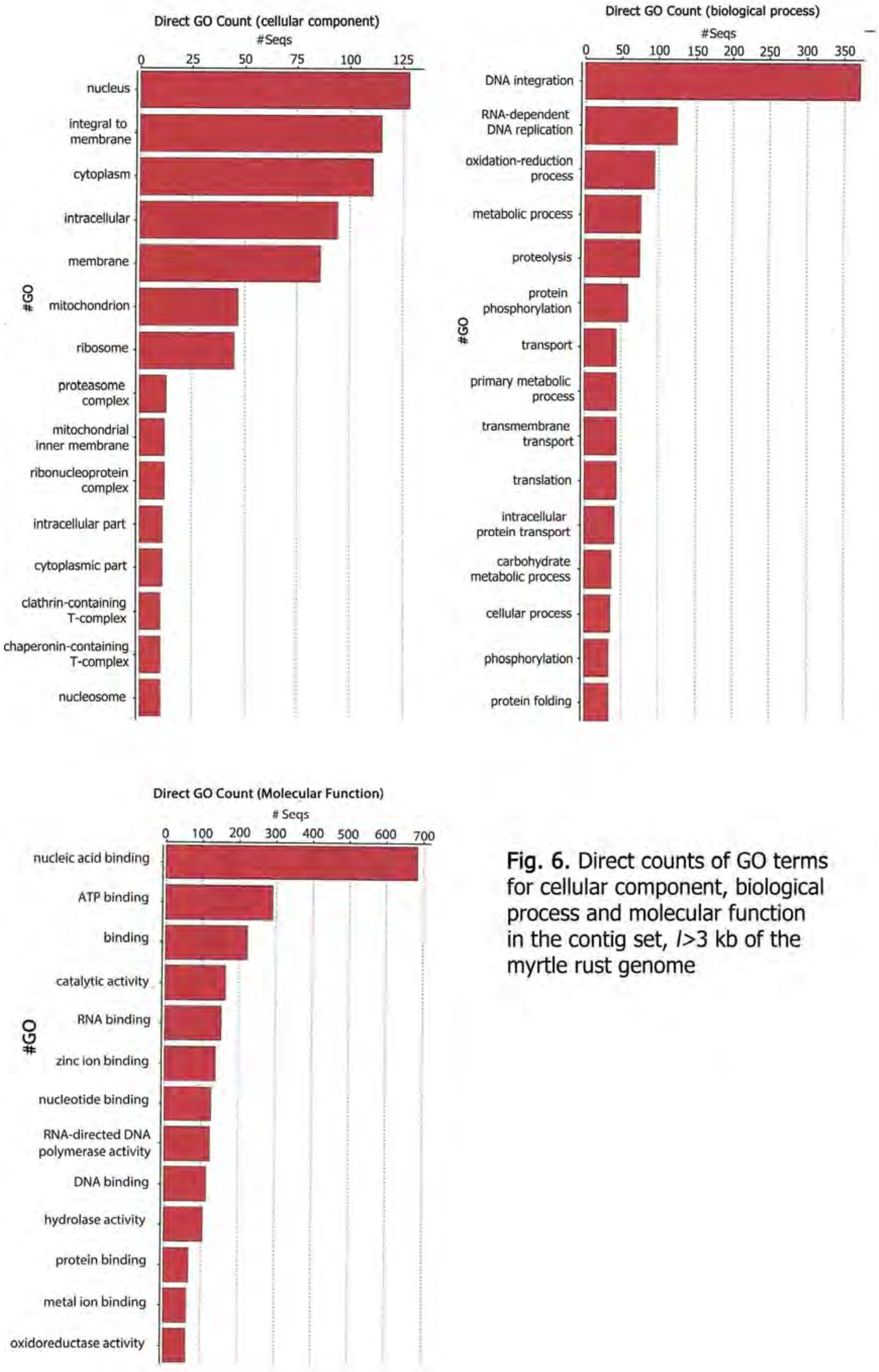

Fig. 6. Direct counts of GO terms for cellular component, biological process and molecular function in the contig set, *l*>3 kb of the myrtle rust genome

4.6. Transposable Elements (TEs)

Transposable elements (TEs) are short, mobile, conserved segments of DNA that can replicate and randomly insert copies within genomes. Eukaryotic TEs are divided into two classes, Class 1 (retrotransposons) and Class II (DNA transposons) (Table 4). The proportion of TEs estimated for the myrtle rust genome was about 27%, with the Class I retrotransposons present in a much higher ratio of about 22% (Table 4).

The DNA transposons (class II elements) are grouped into superfamilies based on sequence similarity of the element-encoded transposase. The proportions of 3 superfamilies have been estimated (Table 4) with the mutator family accounting for ~19% of the DNA transposons (Table 4). The families of more than half of the DNA transposons remain to be accounted.

Table 4. Percentages of different classes of transposable elements for contigs, *l* ≥ 3 kb and *l* ≥ 4 kb assembled using the CLC-Bio Genomics Workbench.

Transposable elements	*l* ≥ 3 kb				*l* ≥ 4 kb			
	# contigs	%	Total length (nucleotide)	%	# contigs	%	Total length (nucleotide)	%
Class1:								
LTR retro-transposons	7,995	21.2	43,373,367	21.4	4,930	21.4	32,851,702	21.6
Class II:								
DNA transposons	1,626	4.3	10,648,124	5.3	1,153	5.1	9,019,956	5.8
[1][	[	[	[	[	[	[	[	[
-mutator	311	19.1	2,040,658	19.1	223	19.3	1,739,036	19.3
-hAT	56	3.4	417,836	3.9	42	3.6	369,208	4.1
-Tc1	29	1.8	161,062	1.5	20	1.7	129,046	1.4
]	]	]	]	]	]	]	]	]

[1][ ] indicates the breakdown of DNA transposons.

4.7. Distribution of contigs coverage

A plot of the distribution of coverage (log$_{10}$ scale) of contigs ≥ 500 bp gave a mean coverage of 10.54 (Fig. 7). Some contigs have very high coverage of more than 10,000 (Table 4). Most of the contigs with abnormally high coverage have transposable elements (LTR-retrotransposons; DNA transposons) annotated on them (Table 5).

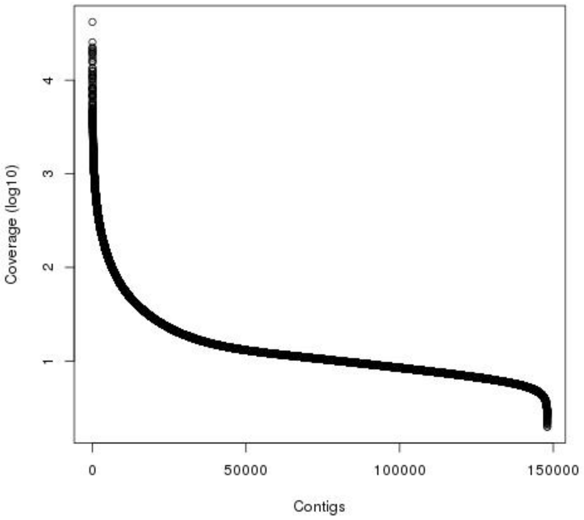

Figure 7. Distribution of coverage ($\log_{10}$ scale) of contigs >500 bp from the CLC-genomics workbench assembled data.

Table 5. Distribution of mean, median, minimum and maximum coverages of transposable elements and hypothetical protein in the contig set, *l* ≥ 3 kb and *l* ≥ 4 kb.

Proteins	**Contig set with *l*≥ 3 kb**				**Contig set with *l*≥ 4 kb**			
	mean	median	min	max	mean	median	min	max
retrotransposable elements	21.5	11.55	4.22	11,286.87	21.4	11.59	4.23	11,286.87
copia polyprotein	16	11.2	4.91	796.6	14.5	11.43	5.02	412.22
gypsy retrotransposon	22.5	12.03	4.62	944.55	18.3	12.00	4.83	538.69
DNA transposons	26.4	11.315	4.61	11,188.22	13.8	11.6	5.23	462.26
hAT	12.7	11.57	6.7	40.34	12.2	11.84	6.7	24.57
DDE	12.4	10.61	5.47	113.58	13.1	11.34	5.8	113.58
mutator	16.3	11.32	4.93	583.75	13.3	11.57	5.54	100.35
Tc1	11.6	11.33	5.76	18.13	11.5	11.22	5.76	18.13
RNase H	18.7	11.5	4.44	896.15	15.4	11.82	5.99	255.31
hypothetical protein	16.7	10.91	3.85	5,704.75	16	11.05	4.33	5,704.75

4.8. Fine Annotation of myrtle rust genes

The boundaries of exons and introns have been determined for 46 protein genes on contigs with length >20,000 nt. The β-tubulin gene was annotated on a contig of length 3,439 bp. The number of exons ranges from 2 to 20 with a mean of 7.6 (Table S4).

4.9. Genes as taxonomic markers and Phylogenetic Analysis

The clathrin-associated protein AP47-coding region was annotated on Contig 5012 (GenBank KF431979). The gene has 9 exons spanning a length of 2,307 nucleotides. A sub-sequence of 307 amino acid residues of the 424 amino acid residues in the myrtle rust protein (GenBank

KF431979) was aligned with orthologous sequences from 10 basidiomycetes; *P. graminis* (2), *M. larici-populina, Mixia osmundae, Rhodosporidium toruloides, Serpula lacrymans, Auricularia delicate, Coniophora puteana, Punctularia strigosozonata* and *Gloeophyllum trabeum* for phylogenetic analysis. The ortholog from *Rhizopus delemar* was used as the outgroup in the analysis. The analysis showed the four rust fungi in a lineage separate from the other basidiomycetes. The rust lineage splits into 2 sub-lineages, one corresponded with *M. larici-populina* and the other with *P. graminis* (Fig. 8). Myrtle rust was clustered on the same sub-lineage as *P. graminis* but on a different branch.

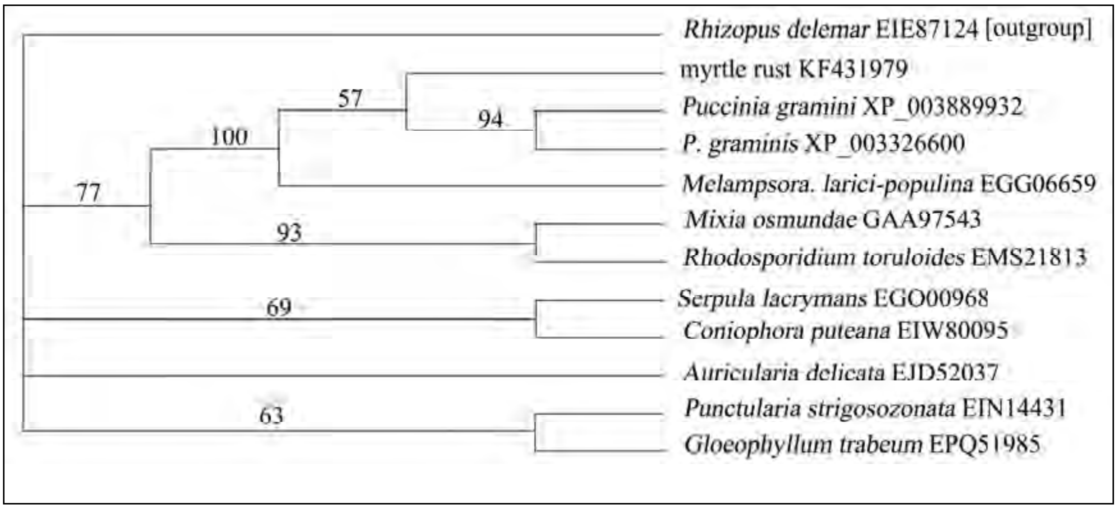
Rhizopus delemar EIE87124 [outgroup]
myrtle rust KF431979
Puccinia gramini XP_003889932
P. graminis XP_003326600
Melampsora. larici-populina EGG06659
Mixia osmundae GAA97543
Rhodosporidium toruloides EMS21813
Serpula lacrymans EGO00968
Coniophora puteana EIW80095
Auricularia delicata EJD52037
Punctularia strigosozonata EIN14431
Gloeophyllum trabeum EPQ51985
57
94
100
77
93
69
63

Figure 8. Phylogeny of the myrtle rust clathrin associated protein with Basidiomycetous orthologs. A consensus tree constructed by the heuristic search option (maximum parsimony criterion) in PAUP4b.10 program using a sub-sequence of 307 amino acids of the clathrin associated protein A47 of myrtle rust isolate (115012_Mr; *P. psidii*) and 10 orthologous sequences from basidiomycetes. The fungus, *Rhizopus delemar* (EIE87124) was used as an outgroup. Only bootstrap values greater than 50% are shown.

The tubulin beta-1 chain was annotated on Contig 103421 (GenBank KF477285). The gene has 14 exons, spanning a length of 2,504 nucleotides. The gene was aligned with the orthologous gene from *P. graminis (*2), *M. larici-populina, Sporisorium reilanum, Ustilago hordei, Pseudozyma flocculosa* and *Penicillium oxalium* (outgroup) for phylogenetic analysis. The results obtained for myrtle rust status were in total agreement with results from the clathrin gene analysis (Fig. 9).

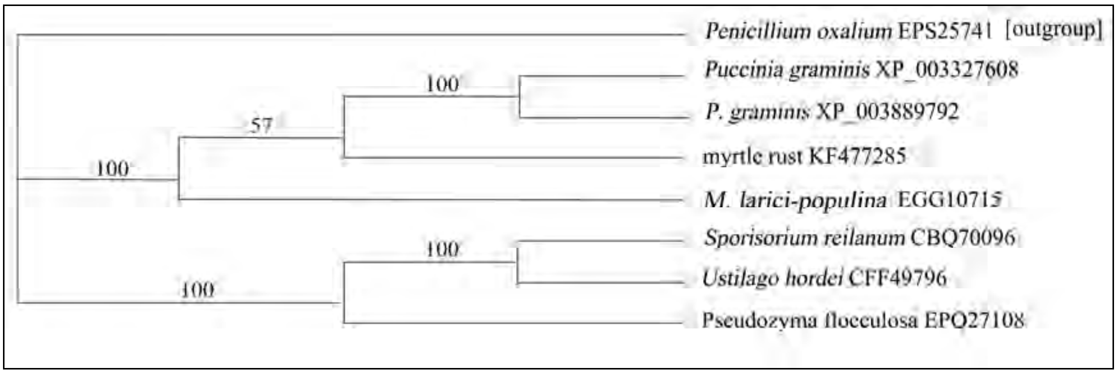
Penicillium oxalium EPS25741 [outgroup]
Puccinia graminis XP_003327608
P. graminis XP_003889792
myrtle rust KF477285
M. larici-populina EGG10715
Sporisorium reilanum CBQ70096
Ustilago hordei CFF49796
Pseudozyma flocculosa EPQ27108
100
57
100
100
100

Figure 9. Phylogeny of the myrtle rust β-tubulin gene with Basidiomycetous orthologs. A consensus tree constructed by the heuristic search option (maximum parsimony criterion) in PAUP4b.10 program using a sequence of 409 amino acids of the β-tubulin gene of myrtle rust isolate (115012_Mr; *P. psidii*) and 6 orthologous sequences from basidiomycetes. The ascomycete, *Penicillium oxalium* (EPS25741) was used as an outgroup. Only bootstrap values greater than 50% are shown.

The complete *COXI* gene has been annotated on Contig 12 (GenBank KF431978). The gene has 11 exons, which with introns span a total length of 18,670 nucleotides. The gene's exons are a total of 1,620 nucleotides, coding 540 amino acids. Due to the large number of introns in the *COXI* gene of basidiomycetes, complete *COXI* gene sequences are limited to date to 5 species.

A short segment (81 amino acids) coding the 5′ end of the *COXI* gene was available in GenBank for many *Puccinia* and *Melampsora* species. These were aligned with orthologous regions of the myrtle rust *COXI* gene and two *Phakopsora* species. Alignment was also made for the corresponding coding DNA sequences of 243 nucleotides.

Analysis using the DNA coding sequences of the partial *COXI* gene fragment with the fungus, *Ustilago maydis* as an outgroup, suggested myrtle rust to belong to the *Pucciniaceae* clade of the *Pucciniales* and in a separate branch from the one for the cereal rust fungi and the *Phakopsora* species (Fig. 10).

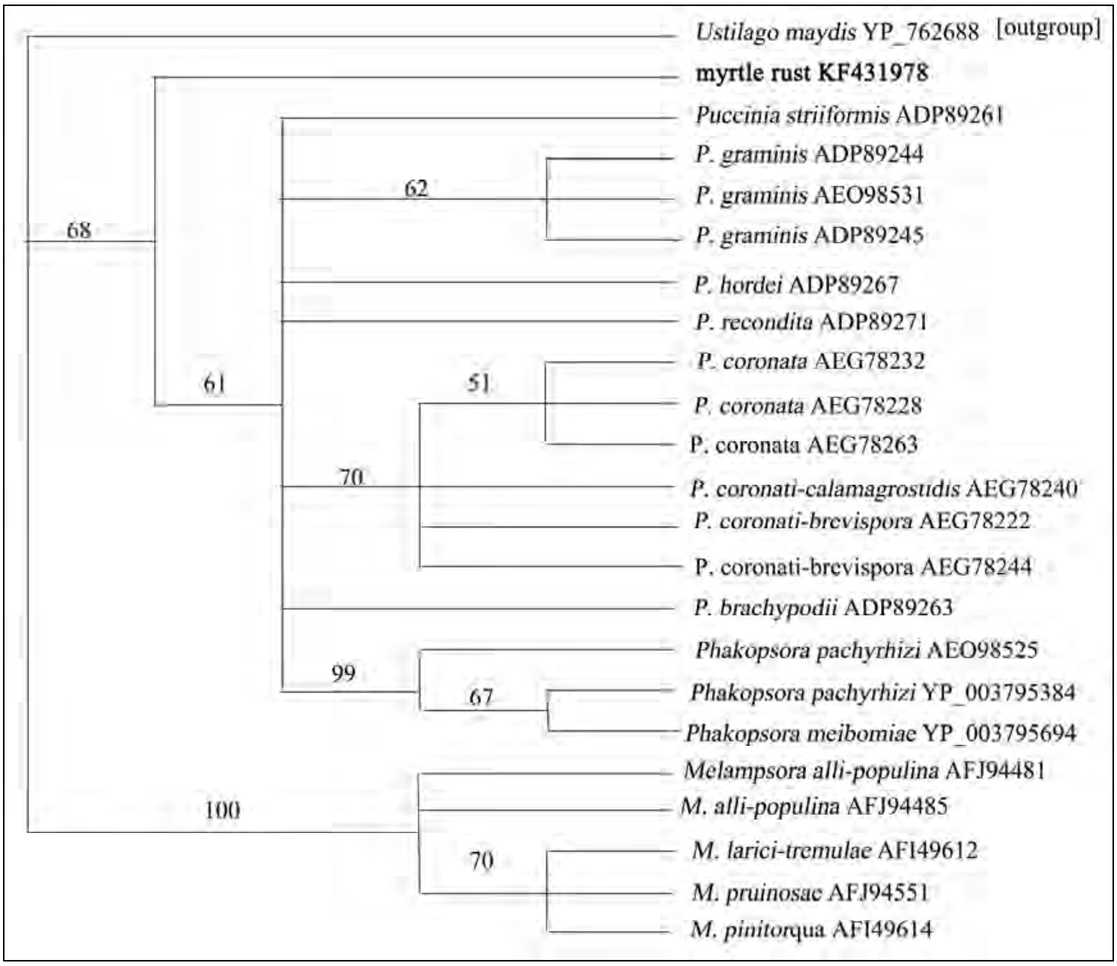
Ustilago maydis YP_762688 [outgroup]
myrtle rust KF431978
Puccinia striiformis ADP89261
P. graminis ADP89244
P. graminis AEO98531
P. graminis ADP89245
P. hordei ADP89267
P. recondita ADP89271
P. coronata AEG78232
P. coronata AEG78228
P. coronata AEG78263
P. coronati-calamagrostidis AEG78240
P. coronati-brevispora AEG78222
P. coronati-brevispora AEG78244
P. brachypodii ADP89263
Phakopsora pachyrhizi AEO98525
Phakopsora pachyrhizi YP_003795384
Phakopsora meibomiae YP_003795694
Melampsora alli-populina AFJ94481
M. alli-populina AFJ94485
M. larici-tremulae AFI49612
M. pruinosae AFJ94551
M. pinitorqua AFI49614
68
62
61
51
70
99
67
100
70

Figure 10. A consensus tree generated using the heuristic search option (maximum parsimony criterion) of PAUP 4b.10 for a short DNA segment of *COXI* gene (5′ end). Only bootstrap values greater than 50% are shown. The coding sequences were obtained from the GenBank accession numbers of the protein sequences which were indicated with the species name.

5. Discussion

5.1. Insights into the genome of myrtle rust

This work has generated 46 million paired end reads of length 250 nucleotides per read on the Illumina MiSeq platform. There is no reference genome for the myrtle rust pathogen. Contigs were assembled *de novo*.

Softwares based on the overlap graph, Readjoiner and SGA used to assemble the data did not perform as well as softwares based on the de Bruijn graph. Of the assemblers based on the de Bruijn graph, the CLC-Genomics Workbench gave the highest N50 value and the longest 'maximum contig length' (Table 2).

The genome size of the myrtle rust pathogen was estimated to be between 103–145 Mb (Table 2). This size is relatively large for fungal genomes (Baker et al. 2008), but comparable to the size of rust genomes reported including *P. triticina* (100-120 Mb) and *P. Striiformis* (110 Mb) (www.broadinstitute.org) and *M. larici-populina* (101 Mb, Duplessis et al. 2011).

The number of proteins in the genome is at least 19,000 (Table 3). This number is in the range of the number of proteins estimated for cereal rust genomes which range from 14,878 for *P. triticina* to 19,542 for *P. striiformis* (http://www.broadinstitute.org/annotation/) and 16,399 and 17,773 for *M. larici-populina* and *P. graminis* f. sp. *tritici* respectively (Duplessis et al. 2011).

The theoretical expected lengths of contigs based on the Lander-Waterman model (Lander and Waterman 1988) for 46 million reads of 250 bp length are predicted to be relatively large (more than 100K bp). The results obtained with various assemblers, however, gave contigs many times shorter (Table 2).

In an effort to understand why the contigs obtained for the myrtle rust pathogen were much shorter than expected, the ABySS software was used to assemble NGS data for the pathogen, *Tilletia indica* obtained using the same methodology to compare the output of genome assemblies with those of the myrtle rust pathogen in this work. The best output for *T. indica* using the ABySS program is with k-mer = 45; giving a N50 value of 16,206 bp and a maximum contig length of 1,141,648 bp (Fig. 11). In contrast, the outputs for myrtle rust using the same program was very poor with N50 values around 500 bp and maximum contig sizes around 13,000 bp regardless of the k-mer value used (Fig. 11).

The maximum contig length from *T. indica* genome is more than a million bp. This is more than 86 times longer than the maximum contig length from the myrtle rust genome. Similarly the N50 value of the *T. indica* genome is more than 20 times longer. These results suggested that the shorter than expected contig length of the myrtle rust genome is due to the inherent genome characteristics of myrtle rust.

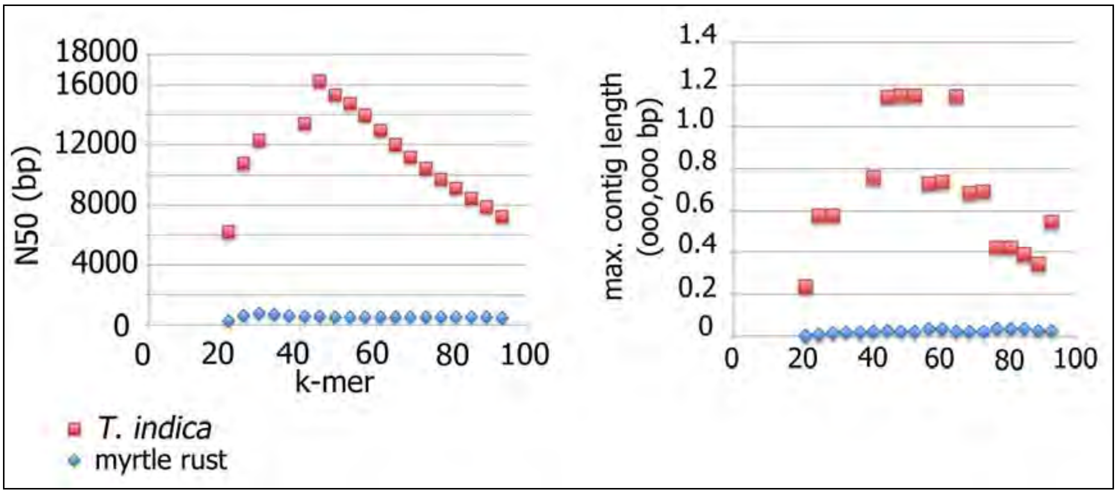

Figure 11. The distribution of N50 values and maximum contig lengths for the assemblies of NGS data of myrtle rust and *T. indica* genomes with different k-mer values on ABySS.

The myrtle rust genome has a large fraction of repeated elements of which a significant percentage comprises transposable elements. The proportion of transposable elements was estimated to be about 27% (Table 4). This is in agreement with recent findings of whole genome sequencing of rust genomes including *P. striiformis* f. sp*. tritic* (Cantu et al. 2011) and the basidiomycete *Laccaria bicolor* (Labbe et al. 2012) which reported a percentage of 17.8% and 24% respectively. Other rust species have been reported to have a much higher percentage of TEs. The rust species, *M. larici-populina* and *P. graminis* f. sp. *tritici* were reported to have 45% of their genomes attributed to TEs (Duplessis et al. 2011).

Eukaryotic TEs are divided into two classes, depending on their mode of transposition: Class I elements (retrotransposons), which mobilize by a 'copy-and-paste' mechanism via a ribonucleic acid (RNA) intermediate, and class II elements (DNA transposons), which move by a 'cut-and-paste' mechanism via a DNA intermediate (Casacuberta and Santiago 2003; Feschotte and Pritham 2007).

The retrotransposons (class I elements) are the most common TE in fungi Muszewska et al. 2007). This is found to be also the case in myrtle rust with the frequency of retrotransposons more than five times higher than DNA transposons (Table 4). Retrotransposons can be classified into two types – LTR retrotransposons and non-LTR retrotransposons (encompassing LINEs and SINEs elements), depending on whether they possess or lack long terminal repeats (LTRs) at both ends. The two main superfamilies of LTR retrotransposons found in fungi are *Gypsy* and *Copia,* which differ in the order of reverse transcriptase (RT), ribonuclease H (RH), and integrase (IN) domains in the virus-like polyprotein (POL;

Gypsy: PR-RT-RH-INT, *Copia*: PR-INT-RT-RH). These two families comprise the largest proportion (~22%) of TEs in the myrtle rust genome (Table 4). The SINEs do not encode a functional reverse transcriptase and rely on other TEs for transposition.

The DNA transposons (class II elements) have terminal inverted repeats (TIRs) or a rolling circle replicon mechanism. They contain a 'DDE motif', which is the active site of the transposase gene. The transposase gene is flanked by a terminal inverted repeat of variable length and catalyzes the "cut and paste" process of the DNA transposon. DNA transposons are grouped into superfamilies (e.g. *hAT*; *Tc1*) based on sequence similarity of the transposase gene.

The superfamilies annotated include *Tc1* (1.7%); *hAT* (3.6%) and *Mutator* (19.3%) and they are signified by the 'DDE' motif. This motif is present in 11 of the 19 currently recognized superfamilies of DNA transposons (http://www.girinst.org/repbase/index.html). The other DDE-domain containing DNA transposons (75.4%) remain to be identified.

The non-DDE domain containing DNA transposons in the myrtle rust genome (potentially 8 superfamilies) if present, have also yet to be accounted. These non-DDE domain containing DNA transposons are probably nonautonomous and use transposase encoded by autonomous elements located elsewhere in the genome. Hence the percentage of DNA transposons will be potentially higher than the estimated 5% (Table 4).

The mean coverage of contigs> 500 bp for the genome assembly of myrtle rust is 10.54 (Fig. 7). In contrast a similar assembly of NGS data from *T. indica* gave a mean coverage of 95 (Tan et al. unpublished data). Most of the contigs with abnormally high coverage have transposable elements annotated on them (Table 5). For instance, the maximum coverages of some of the contigs with LTR-retrotransposons and DNA transposons are greater than 10,000 (Table 5). These repetitive elements were assembled into single contigs due to difficulties in assembling and/or scaffolding with other contigs. These contigs bearing transposons with very high coverage suggest that the real proportion of transposable elements in the genome is higher than the estimated 27%. These TEs have caused serious difficulties in sequence assembly and explain for the large discrepancy between predicted and observed contig sizes for the myrtle rust genome obtained from the *de-novo* assembly.

5.2. Myrtle rust genes as taxonomic markers

Molecular systems for fungal phylogeny have used a multitude of gene regions. The project on 'Assembling the Fungal Tree of Life' (http://www.aftol.org/) includes the 18S rRNA, 28S rRNA, 5.8S rRNA, ITS, *EF1*, *RPB1*, *RPB2, ATP6*. The study of evolutionary relationships in *Pucciniaceae* had reported the use of 18S rRNA (Wingfield et al. 2004), 28S rRNA (Maier et al. 2003) and partial β-tubulin and elongation factor1α gene sequences (Van der Merwe et al. 2007).

DNA barcoding in animals have been based primarily on the mitochondrial *cytochrome c oxidase* subunit 1 (*COXI*) gene. The DNA barcoding system uses a short, standardized gene region of 648 bp as the core barcode region for animals (http://www.dnabarcoding101.org/). The *COXI* gene has been reported to be valuable for the identification of *Penicillium* species (Serfert et al. 2007) and cereal rusts (Liu and Hambleton 2012). The availability of sequences for a short segment of the mitochondrial *COXI* gene for a large number of rust species facilitated the use of the short *COXI* gene segment for the analysis of evolutionary relationships of myrtle rust with other rust species.

Most eukaryotic genes contain introns and the boundaries between exons and introns in the genes must be determined before they can be used appropriately for phylogenetic analysis. The clathrin-associated protein, beta-tubulin gene and the *COXI* gene of the myrtle rust has 9, 14 and 11 exons respectively (Table S4).

Based on the availability of protein orthologs from species in *Puccinia* and *Melampsora* and other basidiomycetes, the protein sequence of the β-tubulin gene and the clathrin associated protein A47 were used to construct the phylogeny of myrtle rust with respect to other fungi particularly the rust fungi and other basidiomycetes.

The number of rust sequences was limited for both analyses, with only sequences of *M. larici-populina* and *P. graminis* available as representatives of *Melampsoraceae* and *Pucciniaceae* respectively of the *Pucciniales*. Myrtle rust was clustered with *Pucciniaceae* in both analyses (Fig. 8, 9).

The β-tubulin gene sequence of myrtle rust (KF477285) is the same variant as *P. graminis* (XP_003327608, XP_003889792) and *M. larici-populina* (EGG10715) but these sequences apparently represent a different variant to the partial β-tubulin gene sequences used in the study of evolutionary relationships in *Puccinia* and *Uromyces* (van der Merwe et al. 2007). Hence, the myrtle rust sequence could not be incorporated for a more in-depth analysis.

The *COXI* gene in basidiomycetes can be interrupted by multiple large introns at variable locations. The *COXI* gene in myrtle rust is interrupted by 10 introns of length ranging from 1,032 to 2,562 bp (KF431978). The DNA fragment from myrtle rust genome used in the phylogenetic analysis (Fig. 10) comprises exon 1 and exon 2 separated by an intron of 2,552 nucleotides. The intron was excluded from the genomic sequence to give the coding sequence for phylogenetic analysis with other rust species (Fig. 10).

The *COXI* segment used in this study was short compared to the 648 bp sequence used in animal barcoding. However, the DNA analysis (Fig. 10) had resolved the rust species used in this study into two distinct, major clades, *Pucciniaceae* and *Melampsoraceae* in agreement with previous studies (Maier et al. 2003, 2007; Wingfield et al. 2004; Van der Merwe et al. 2007). The *Phakopsora* species were included in the same clade as the *Puccinia* species and this corroborated the outcome of analysis using the 18S rRNA gene (Wingfield et al. 2004). This study suggested that myrtle rust belongs to the major clade, *Pucciniaceae* of *Pucciniales*, but on a separate taxonomic branch from the one for the cereal rust fungi and *Phakopsora* species (Fig. 10).

The three taxonomic markers in this study have indicated myrtle rust to be in the *Pucciniaceae* clade. However, Van der Merwe et al. 2008 suggested that *P. psidii* and *Phakopsora pachyrhizi* were excluded from *Pucciniaceae* in their analysis using partial β-tubulin sequences. There is thus a need to use more sequences of rust on *Myrtaceae* (including guava and eucalyptus rust) for further confirmation.

5.3. Implications and Recommendations

This study has found that the myrtle rust genome has about 27% of transposable elements. They are major players in the generation of genetic variability in the pathogen’s adaptation to the environment and new hosts.

The very wide host range of myrtle rust in so many genera would imply that the myrtle rust pathogen, like the cereal rusts, is potentially a complex with taxonomic levels varying from races to species. Simpson et al (2006) documented 8 taxa of rusts on *Myrtaceae* with different host ranges and specificity. A review had reported 27 synonyms for *P. psidii* (Glen et al. 2007), all of which were found on *Myrtaceae* hosts. All this suggests the myrtle rust genome is continuously evolving to overcome host resistance, and this co-evolution with their hosts will lead to the generation of a rust complex with different host specific genotypes.

Rusts in *Pucciniaceae* have been documented to co-evolve with their angiospermous hosts to enable them to jump across hosts of different species and genera (van der Merwe et al. 2008). Eucalyptus rust is apparently a specialized genotype evolved from a rust on *Myrtaceae* in South America to enable it to 'host jump' and infect *Eucalyptus*, an introduced *Myrtaceae* species to Brazil (Ferreira 1983) and hence its name. Isolates from guava did not infect *Eucalyptus* and vice versa (Ferreira 1983) indicating the evolution of host-specific genotypes.

The 'plasticity' of rusts to adapt to new hosts poses a huge threat on commercially important Australian *Eucalyptus* and some endangered *Myrtaceae* genera in the Australian flora. It is thus very important to have an on-going, long term program to understand and manage the pathogen. Critical areas would include the study of the genetic diversity of this pathogen and their associated host range and to monitor the evolutionary changes occurring with respect to changes in pathogenic potential and host specificity.

Further research on genetic diversity studies in myrtle rust should include the closely related guava rust and eucalyptus rust where to-date no genome sequence data is available for comparison. This will enable accurate diagnosis and identification of the different genotypes in the myrtle rust complex for a more efficient implementation of management strategies and breeding programs for resistant germplasm. Accurate identification of the myrtle rust isolates is also crucial for the regulation of movement of *Myrtaceae* materials, particularly *Eucalyptus* between geographical areas and countries.

6. References

Baker SE, Thykaer J, Adney WS, Brettin T, Brockman FJ et al (2008) Fungal genome sequencing and bioenergy. Fungal Biology Reviews 22:1–5

Begerow D, John B, Oberwinker F (2004) Evolutionary relationships among beta-tubulin gene sequences of basidiomycetous fungi. Mycological Research 108(11): 1257-1263

Cantu D, Govindarajulu M, Kozik A, Wang M, Chen X, et al. (2011) Next Generation Sequencing Provides Rapid Access to the Genome of

Puccinia striiformis f. sp. tritici, the Causal Agent of Wheat Stripe Rust. PLoS ONE 6(8): e24230. doi:10.1371/journal.pone.0024230

Carnegie A, Lidbetter JR (2012) Rapidly expanding host range for *Puccinia psidii* sensu lato in Australia. Australasian Plant Pathology 41:13–29

Carnegie AJ, Cooper K (2011) Emergency response to the incursion of an exotic myrtaceous rust in Australia. Australasian Plant Pathology 40:346–359

Carnegie AJ, Lidbetter JR, Walker J, Horwood MA, Tesoriero L et al (2010) *Uredo rangelii*, a taxon in the guava rust complex, newly recorded on *Myrtaceae* in Australia. Australasian Plant Pathology 39:463–466

Casacuberta JM, Santiago N (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene 311:1–175

Coutinho TA, Wingfield MJ, Alfenas AC, Crous PW (1998) Eucalyptus rust: a disease with the potential for serious international implications. Plant Disease 82:819–825

Desper R, Gascuel O (2004) Theoretical Foundation of the Balanced Minimum Evolution Method of Phylogenetic Inference and Its Relationship to Weighted Least-Squares Tree Fitting. Mol Biol Evol 21(3):587–598

Duplessis S, Cuomob CA, Lin Y, Aerts A, Tisseranta E et al (2011) Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc Natl Acad Sci USA 108(22):9166–9171

Ferreira FA(1983) Eucalyptus rust. Revista Arvore 7:91–109

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes [Review] Annual Review of Genetics. 41:331–368

Furnier GR, Stolz AM, Mustaphi RM, Ostry ME (1999) Genetic evidence that butternut canker was recently introduced into North America. Can J Botany 77(6):783–785

Glen M, Alfenas AC, Zauza EAV, Wingfield MJ, Mohammed C (2007) *Puccinia psidii*: a threat to the Australian environment and economy –a review. Australasian Plant Pathology 36:1-16

Götz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the blast2go suite. Nucleic Acids Res 36(10):3420–3435

Gonnella G, Kurtz S (2012) Readjoiner: a fast and memory efficient string graph-based sequence assembler. BMC Bioinformatics 13:82

Junghans DT, Alfenas AC, Brommonschenkel SH, Oda S, Mello EJ, Grattapaglia D (2003) Resistance to rust (*Puccinia psidii* Winter) in *Eucalyptus*: mode of inheritance and mapping of a major gene with RAPD markers. Theoretical & Applied Genetics 108:175–180

Labbe ́ J, Murat C, Morin E, Tuskan GA, Le Tacon F, et al. (2012) Characterization of Transposable Elements in the Ectomycorrhizal Fungus *Laccaria bicolor*. PLoS ONE 7(8): e40197.

Lander ES,Waterman MS (1988) Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics 2:231–239

Liu M, Hambleton S (2012) *Puccinia chunjii*, a close relative of the cereal stem rusts revealed by molecular phylogeny and morphological study. Mycologia 104(5):1056–1067

Luo R, Liu B, Xie Y, Li Z, Huang W et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18

Maier W, Begerow D, Weiβ M, Oberwinkler F (2003) Phylogeny of the rust fungi: an approach using the nuclear large subunit ribosomal DNA sequences. Canadian Journal of Botany 81:12–23

Maier W, Wingfield BD, Mennicken M, Wingfield MJ (2007) Polyphyly and two emerging lineages in the rust genera *Puccinia* and *Uromyces*. Mycol Res 111:(2)176–185

Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315–327

Muszewska A, Hoffman-Sommer M, Grynberg M (2011) LTR Retrotransposons in Fungi. PLoS ONE 6(12): e29425.

Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci USA 98:9748–9753

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Schatz MC, Delcher AL, Salzberg SL (2010) Assembly of large genomes using second-generation sequencing. Genome Res 20:1165–1173

Seifert KA, Samson RA, deWaard JR, Houbraken J, Levesque CA, et al. (2007) Prospects for fungus identification using CO1 DNA barcodes, with *Penicillium* as a test case. Proc Natl Acad Sci USA 104:3901–3906

Simpson JA, Thomas K, Grgurinovic CA (2006) Uredinales species pathogenic on species of *Myrtaceae*. Australasian Plant Pathology 35:549–562

Simpson JT, Durbin R (2012) Efficient *de novo* assembly of large genomes using compressed data structures. Genome Res 22(3):549–556

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) Abyss: A parallel assembler for short read sequence data. Genome Res 19:1117–1123

Tan MK, Niessen LM (2003) Analysis of rDNA ITS sequences to determine genetic relationships among, and provide a basis for simplified diagnosis of, Fusarium species causing crown rot and head blight of cereals. Mycol Res 107(7):811–821

van der Merwe M, Ericson L, Walker J, Thrall PH, Burdon JJ (2007) Evolutionary relationships among species of *Puccinia* and *Uromyces* (Pucciniaceae, Uredinales) inferred from partial protein coding gene phylogenies. Mycol Res 111(2):163–175

van der Merwe MM, Walker J, Ericson L, Burdon JJ (2008) Coevolution with higher taxonomic host groups within the *Puccinia*/*Uromyces* rust lineage obscured by host jumps Mycol Res 112(12):1387–1408

Wingfield BD, Ericson L, Szaro T, Burdon JJ (2004) Phylogenetic patterns in the uredinales. Australasian Plant Pathology 33:327–335

Winter G (1884) Repertorium. Rabenhorstii fungi europaei et extraeuraopaei. Centuria XXXI et XXXII. Hedwigia 23:164–175

7. Supporting Information

Table S1: Summary of assembly results using Readjoiner (with different overlap lengths) on trimmed NGS data of myrtle rust

	l≥200				*l*≥750			
Overlap length, *ol*	N25	N50	Total # of sequences	Total length of sequences (bp)	N25	N50	Total # of sequences	Total length of sequences (bp)
40	534	434	4361413	1846223949	954	849	60789	53339207
50	536	438	4683397	2003394714	963	853	82966	73237092
60	538	440	4867176	2097657721	970	857	102242	90634325
70	538	439	4963090	2147976637	973	859	116914	103828908
80	538	438	5001053	2167240038	973	858	127058	112780714
90	536	436	5002593	2163677734	969	857	131152	116228519
100	532	432	4985177	2144451081	965	855	129227	114152489

l≥200– contigs larger than threshold (200bp); *l*≥750– contigs larger than threshold (750bp);

Table S2: Summary of assembly results using different k-mer on the ABySS software.

K-mer	n	*l*≥200	n:N50	min	N80	N50	N20	max	sum
45	15.49e6	759088	155393	200	275	558	1526	25668	362e6
49	13.73e6	890044	188245	200	274	546	1434	20557	419.1e6
53	12.2e6	1017635	220121	200	274	541	1362	20561	475.1e6
57	10.86e6	1137295	250176	200	274	540	1314	33667	528.8e6
61	9729172	1249840	278077	200	274	540	1280	33659	579.2e6
65	8739352	1352915	304736	200	274	539	1245	20332	624.9e6
69	7857677	1449296	329775	200	273	538	1218	22105	666.5e6
73	7108705	1537936	353548	200	272	535	1191	22047	703.2e6
77	6446782	1615042	374947	200	272	532	1165	35730	734.2e6
81	5848562	1680314	393701	200	271	528	1142	35730	759.3e6
85	5333221	1737654	411596	200	270	522	1116	35727	779.5e6
89	4870808	1786328	427232	200	269	515	1091	25527	794.7e6

n — total contigs in the assembly; *l*≥200— contigs larger than threshold (200bp);
n:N50 — number of contigs contained in the N50 set; min — smallest contig;
N50 — N50 contig length; max — largest contig; sum — sum of contig lengths

Table S3: Summary of assembly results using different k-mer on the CLC-Bio Genomics Workbench (excluding scaffolded regions)

k-mer	N75	N50	N25	Min	Max	Median	Number of contigs	Total length
32	1598	2639	4532	120	39311	2148	163338	350,913,382
40	**1739**	**3059**	**5555**	**699**	**47817**	**2402**	**169920**	**386,562,065**
45	1739	3043	5506	732	42282	2394	171922	411,641,477
50	1722	2967	5372	655	42350	2360	183767	433,610,821
64	1631	2665	4617	249	50064	2208	220120	486,116,824

Table S4: List of fully annotated genes of the myrtle rust pathogen with their corresponding GenBank accession numbers

Gene	# exons	Contig length	GenBank Accession #
AarF; Pkc_like	5	22551	KF431993
AdoMet Mtases	6	22485	KF431980
ATP12	7	22485	KF431980
ATP_sub_h	2	20659	KF431974
C2_RasGAP	3	33008	KF431975
clathrin associated protein	9	23635	KF431979
Cpn60_TCP1	8	20034	KF431976
DNA repair protein (rad1)	14	32354	KF431977
DnaJ	7	21047	KF431981
DSPc	6	20930	KF431982
eukaryotic translation initiation factor 2 subunit	7	26736	KF431983
farnesyl-diphosphate farnesyltransferase	8	24879	KF431984
FAT; TRRAP; PI3Kc	14	31264	KF431988
GIT_SHD	10	21927	KF431985
Glyco_hydro_2_C	12	28805	KF431986
Glyco_transf_25	2	22428	KF431989
glycoside hydrolase family 92 protein	20	20606	KF431987
Heterokaryon incompatibility protein Het-C	17	26736	KF431990
IbpA_ACD_LpsHSP_like	3	33590	KF431991
MBOAT_2	3	21660	KF431992
nadF	5	21393	KF431994
Patatins and Phospholipases	6	31522	KF431995
Pectate_lyase_3	2	22483	KF431996
Peptidase_C14 (Caspase domain; pfam00656)	11	24405	KF431997
Peptidase_M14NE-CP-C_like	6	29845	KF431998
peptidase_M17	9	25580	KF431999
peptidylprolyl isomerase	4	24879	KF432000
Phox homology (PX) domain protein (COG5391)	8	20306	KF432001
rab family protein	4	22719	KF432002
Ras-like protein Rab7	6	30285	KF432003
Ribosomal_P0_like	5	37821	KF432004
RINT-1 _ TIP-1 family; pfam04437	9	20112	KF432005
SCAMP family; pfam04144	5	24168	KF432006
SecE	4	20357	KF432007
Sen15	4	22496	KF432008
SF3b1_HSH155	13	21811	KF432009
Sfi1	11	23803	KF432010
small nuclear ribonucleoprotein D3	4	24696	KF432011
SPX_CitT_SLC13 permease	5	20601	KF432012
Sun_AdoMet_MTases	10	20962	KF432013
TFIIE beta winged helix	7	23401	KF432014
ubiquitin thiolesterase	9	20356	KF432015
Ubox_RING_cyclophilin_RING	11	29845	KF432016
Uncharacterized conserved protein_COG0397	5	20101	KF432017
WD40_Peptidase_C19_UCH_1_PAN2_exo	10	22935	KF432018
COXI gene	11	38639	KF431978
tubulin beta-1 chain	14	3438	KF477285

8. Abbreviations/Glossary

ABBREVIATION	FULL TITLE
ATP binding	GO:0005524 (molecular function); Interacting selectively and non-covalently with ATP, adenosine 5'-triphosphate, a universally important coenzyme and enzyme regulator.
Binding	GO:0005488 (molecular function); The selective, non-covalent, often stoichiometric, interaction of a molecule with one or more specific sites on another molecule.
Cytoplasmic part	GO:0044444 (cellular component); Any constituent part of the cytoplasm, all of the contents of a cell excluding the plasma membrane and nucleus, but including other subcellular structures.
DNA integration	GO:0015074 (Biological Process); The process in which a segment of DNA is incorporated into another, usually larger, DNA molecule such as a chromosome.
E-value	a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. The lower the value, the more "significant" the match is.
exon	Coding sequence of DNA present in mature messenger RNA
GO	Gene Ontology (The GO project[http://www.geneontology.org/] provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data.
Integral to membrane	GO:0016021 (cellular component); Penetrating at least one phospholipid bilayer of a membrane. When used to describe a protein, indicates that all or part of the peptide sequence is embedded in the membrane.
Intracellular part	GO:0044424 (cellular component); Any constituent part of the living contents of a cell; the matter contained within (but not including) the plasma membrane, usually taken to exclude large vacuoles and masses of secretory or ingested material. In eukaryotes it includes the nucleus and cytoplasm.
intron	Non-coding sequence of DNA removed from mature messenger RNA prior to translation.
l	Contig length
N50	An N50 contig size of N means that 50% of the assembled bases are contained in contigs of length N or larger. N50 sizes are often used as a measure of assembly quality because they capture how much of the genome is covered by relatively large contigs.
NGS	Next Generation Sequencing
Nucleic acid binding	GO:0003676 (molecular function); Interacting selectively and non-covalently with any nucleic acid.
Nucleotide binding	GO:0000166 (molecular function); Interacting selectively and non-covalently with a nucleotide, any compound consisting of a nucleoside that is esterified with (ortho)phosphate or an oligophosphate at any hydroxyl group on the ribose or deoxyribose.
orthologs	Genes in different species that originated by vertical descent from a single gene of the last common ancestor
RNA binding	GO:0003723 (molecular function); Interacting selectively and non-covalently with an RNA molecule or a portion thereof.
RNA-dependent DNA replication	GO:0006278 (Biological Process); A DNA replication process that uses RNA as a template for RNA-dependent DNA polymerases (e.g. reverse transcriptase) that synthesize the new strands.
UnitProtKB	is a protein knowledgebase (http://www.uniprot.org/) which consists of two sections; Swiss-Prot and TrEMBL.